Optimal Estimation and Prediction for Dense Signals in High-Dimensional Linear Models

Lee Dicker

Department of Statistics and Biostatistics
Rutgers University
501 Hill Center, 110 Frelinghuysen Road
Piscataway, NJ 08854
e-mail: ldicker@stat.rutgers.edu

Abstract: Estimation and prediction problems for dense signals are often framed in terms of minimax problems over highly symmetric parameter spaces. In this paper, we study minimax problems over ℓ^2-balls for high-dimensional linear models with Gaussian predictors. We obtain sharp asymptotics for the minimax risk that are applicable in any asymptotic setting where the number of predictors diverges and prove that ridge regression is asymptotically minimax. Adaptive asymptotic minimax ridge estimators are also identified. Orthogonal invariance is heavily exploited throughout the paper and, beyond serving as a technical tool, provides additional insight into the problems considered here. Most of our results follow from an apparently novel analysis of an equivalent non-Gaussian sequence model with orthogonally invariant errors. As with many dense estimation and prediction problems, the minimax risk studied here has rate d/n, where d is the number of predictors and n is the number of observations; however, when $d \asymp n$ the minimax risk is influenced by the spectral distribution of the predictors and is notably different from the linear minimax risk for the Gaussian sequence model (Pinsker, 1980) that often appears in other dense estimation and prediction problems.

AMS 2000 subject classifications: Primary 62J05; secondary 62C20.
Keywords and phrases: adaptive estimation, asymptotic minimax, non-Gaussian sequence model, oracle estimators, ridge regression.

1. Introduction

This paper is about estimation and prediction problems involving non-sparse (or "dense") signals in high-dimensional linear models. By contrast, a great deal of recent research into high-dimensional linear models has focused on sparsity. Though there are many notions of sparsity (e.g. ℓ^p-sparsity (Abramovich et al., 2006)), a vector $\boldsymbol{\beta} \in \mathbb{R}^d$ is typically considered to be sparse if many of its coordinates are very close to 0. Perhaps one of the general principals that has emerged from the literature on sparse high-dimensional linear models may be summarized as follows: if the parameter of interest is sparse, then this can often be leveraged to develop methods that perform very well, even when the number of predictors is much larger than the number of observations. Indeed, powerful theoretical performance guarantees are available for many methods developed under this paradigm, provided the parameter of interest is sparse (Bickel et al., 2009; Bunea et al., 2007; Candès and Tao, 2007; Fan and Lv, 2011; Rigollet and Tsybakov, 2011; Zhang, 2010). Furthermore, in many applications – especially in engineering and signal processing – sparsity assumptions have been repeatedly validated (Donoho, 1995; Duarte et al., 2008; Erlich et al., 2010; Lustig et al., 2007; Wright et al., 2008). However, there is less certainty about the manifestations of sparsity in other important applications where high-dimensional data is abundant. For example, several recent papers have questioned the degree of sparsity in modern genomic datasets (see, for instance, (Hall et al., 2009), and the references contained therein – including (Goldstein, 2009; Hirschhorn, 2009; Kraft and Hunter, 2009) – and, more recently, (Bansal et al., 2010; Manolio, 2010)). In situations like these, sparse methods may be sub-optimal and methods designed for dense problems may be more appropriate.

Let d and n denote the number of predictors and observations, respectively, in a linear regression problem. In dense estimation and prediction problems, where the parameter of interest is not assumed to be sparse, $d/n \to 0$ is often required to ensure consistency. Indeed, this is the case for the problems considered in this paper. In this sense, dense problems are more challenging than sparse problems, where consistency may be possible when $d/n \to \infty$. This lends credence to Friedman et al.'s (2004) "bet on sparsity" principle for high-dimensional data analysis:

> Use a procedure that does well in sparse problems, since no procedure does well in dense problems.

The "bet on sparsity" principle has proven to be very useful, especially in applications where sparsity prevails, and it may help to explain some of the recent emphasis on understanding sparse problems. However, the emergence of important problems in high-dimensional data analysis where the role of sparsity is less clear highlights the importance of characterizing and thoroughly understanding dense problems in high-dimensional data analysis. This paper addresses some of these problems.

Minimax problems over highly symmetric parameter spaces have often been equated with dense estimation problems in many statistical settings (Donoho and Johnstone, 1994; Johnstone, 2011). In this paper, we study the minimax risk over ℓ^2-balls for high-dimensional linear models with Gaussian predictors. We identify several informative, asymptotically equivalent formulations of the problem and provide a complete asymptotic solution when the number of predictors d grows large. In particular, we obtain sharp asymptotics for the minimax risk that are applicable in any asymptotic setting where $d \to \infty$ and we show that ridge regression estimators (Hoerl and Kennard, 1970; Tikhonov, 1943) are asymptotically minimax. Adaptive asymptotic minimax ridge estimators are also discussed. Our results follow from carefully analyzing an equivalent non-Gaussian sequence model with orthogonally invariant errors and the novel use of two classical tools – Brown's identity (Brown, 1971) and Stam's inequality (Stam, 1959) – to relate this sequence model to the Gaussian sequence model with iid errors. The results in this paper share some similarities with those found in (Goldenshluger and Tsybakov, 2001, 2003), which address minimax prediction over ℓ^2-ellipsoids. However, the implications of our results and the methods used to prove them differ substantially from Goldenshluger and Tsybakov's (this is discussed in more detail in Sections 2.2-2.3 below).

2. Background and preliminaries

2.1. Statistical setting

Suppose that the observed data consists of outcomes $y_1, ..., y_n \in \mathbb{R}$ and d-dimensional predictors $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$. The outcomes and predictors follow a linear model and are related via the equation

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, ..., n, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^T \in \mathbb{R}^d$ is an unknown parameter vector (also referred to as "the signal") and $\epsilon_1, ..., \epsilon_n$ are unobserved errors. To simplify notation, let $\mathbf{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$, $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$. Then (1) may be rewritten as $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. In many high-dimensional settings it is natural to consider the predictors $\mathbf{x}_i$ to be random. In this paper, we assume that

$$\mathbf{x}_1, ..., \mathbf{x}_n \overset{\text{iid}}{\sim} N(0, I) \text{ and } \epsilon_1, ..., \epsilon_n \overset{\text{iid}}{\sim} N(0, 1) \tag{2}$$

are independent, where $I = I_d$ is the $d \times d$ identity matrix. These distributional assumptions impose significant additional structure on the linear model (1). However, similar models have been studied previously (Baranchik, 1973; Breiman and Freedman, 1983; Brown, 1990; Leeb, 2009; Stein, 1960) and we believe that the insights imparted by the resulting simplifications are worthwhile. For the results in this paper, perhaps the most noteworthy simplifying consequence of the normality assumption (2) is that the distributions of $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are invariant under orthogonal transformations.

We point out that the assumption $E(\mathbf{x}_i) = 0$ (which is implicit in (2)) is not particularly limiting: if $E(\mathbf{x}_i) \neq 0$, then we can reduce to the mean 0 case by centering and decorrelating the data. If $\text{Var}(\epsilon_i) = \sigma^2 \neq 1$ and σ^2 is known, then this can easily be reduced to the case where $\text{Var}(\epsilon_i) = 1$. If σ^2 is unknown and $d < n$, then σ^2 can be effectively estimated and one can reduce to the case where $\text{Var}(\epsilon_i) = 1$ (Dicker, 2012). We conjecture that σ^2 can be effectively estimated when $d > n$, provided $\sup d/n < \infty$ (for sparse $\boldsymbol{\beta}$, Sun and Zhang (2011) and Fan et al. (2012) have shown that σ^2 can be estimated when $d \gg n$). Dicker (2012) has discussed the implications if $\mathrm{Cov}(\mathbf{x}_i) = \Sigma \neq I$. Essentially, when the emphasis is prediction and non-sparse signals, if a norm-consistent estimator for $\mathrm{Cov}(\mathbf{x}_i) = \Sigma$ is available, then it is possible to reduce to the case where $\mathrm{Cov}(\mathbf{x}_i) = I$; if a norm-consistent estimator is not available, then limitations entail, however, these limitations may not be overly restrictive (this is discussed further in Section 3.2 below).

Let $||\cdot|| = ||\cdot||_2$ denote the ℓ^2-norm. In this paper we study the performance of estimators $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ with respect to the risk function

$$R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = R_{d,n}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = E_{\boldsymbol{\beta}}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2, \tag{3}$$

where the expectation is taken over $(\boldsymbol{\epsilon}, X)$ and the subscript $\boldsymbol{\beta}$ in $E_{\boldsymbol{\beta}}$ indicates that $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ (below, for expectations that do not involve $\mathbf{y}$, we will often omit this subscript). We emphasize that the expectation in (3) is taken over the predictors X as well as the errors $\boldsymbol{\epsilon}$, i.e. it is *not* conditional on X. The risk $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ is a measure of estimation error. However, it can also be interpreted as the unconditional out-of-sample prediction error (predictive risk) associated with the estimator $\hat{\boldsymbol{\beta}}$ (Breiman and Freedman, 1983; Leeb, 2009; Stein, 1960).

2.2. Dense signals, sparse signals, and ellipsoids

Let $B(c) = B_d(c) = \{\boldsymbol{\beta} \in \mathbb{R}^d;\ ||\boldsymbol{\beta}|| \leq c\}$ denote the ℓ^2-ball of radius $c \geq 0$. Though a given signal $\boldsymbol{\beta} \in \mathbb{R}^d$ is often considered to be dense if it has many nonzero entries, when studying broader properties of dense signals and dense estimators it is common to consider minimax problems over highly symmetric, convex (or loss-convex (Donoho and Johnstone, 1994)) parameter spaces. Following this approach, one of the primary quantities that we use as a benchmark for evaluating estimators and determining performance limits in dense estimation problems is the minimax risk over $B(c)$:

$$R^{(b)}(c) = R^{(b)}_{d,n}(c) = \inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in B(c)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}). \tag{4}$$

The infimum on the right-hand side in (4) is taken over all measurable estimators $\hat{\boldsymbol{\beta}}$ and the superscript "b" in $R^{(b)}(c)$ indicates that the relevant parameter space is the ℓ^2-ball.

A basic consequence of the results in this paper is $R^{(b)}(c) \asymp d/n$. Thus, one must have $d/n \to 0$ in order to ensure consistent estimation over $B(c)$. This is a well-known feature of dense estimation problems and, as mentioned in Section 1, contrasts with many results on sparse estimation that imply $\boldsymbol{\beta}$ may be consistently estimated when $d/n \to \infty$. However, the sparsity conditions on $\boldsymbol{\beta}$ that are required for these results may not hold in general and our motivating interest lies precisely in such situations. In this paper we derive sharp asymptotics for $R^{(b)}(c)$ and related quantities in settings where $d/n \to 0$, $d/n \to \rho \in (0, \infty)$, and $d/n \to \infty$

(we assume that $d \to \infty$ throughout). Though consistent estimation is only guaranteed when $d/n \to 0$, there are important situations where one might hope to analyze high-dimensional datasets with d/n substantially larger than 0, even if there is little reason to believe that sparsity assumptions are valid. The results in this paper provide detailed information that may be useful in situations like these.

In addition to sparse estimation problems, minimax rates faster than d/n have also been obtained for minimax problems over ℓ^2-ellipsoids, which have been studied extensively in situations similar to those considered here (Cavalier and Tsybakov, 2002; Goldenshluger and Tsybakov, 2001, 2003; Pinsker, 1980). Much of this work has been motivated by problems in nonparametric function estimation. The results in this paper are related to many of these existing results, however, there are important differences – both in their implications and the techniques used to prove them. Goldenshluger and Tsybakov's (2001, 2003) work may be most closely related to ours. Define the ℓ^2-ellipsoid $B(c, \boldsymbol{\alpha}) = \{\boldsymbol{\beta} \in \mathbb{R}^d;\ \sum_{i=1}^n \alpha_i \beta_i^2 \leq c^2\}$, with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_d)^T \in \mathbb{R}^d$, $0 \leq \alpha_1 \leq \cdots \leq \alpha_d$. Goldenshluger and Tsybakov studied minimax problems over ℓ^2-ellipsoids for a linear model with random predictors similar to the model considered here (in fact, Goldenshluger and Tsybakov's results apply to infinite-dimensional non-Gaussian $\mathbf{x}_i$, though $\mathbf{x}_i$ are required to have Gaussian tails and independent coordinates). They identified asymptotically minimax estimators over $B(c, \boldsymbol{\alpha})$ and adaptive asymptotically minimax estimators and showed that the minimax rate may be substantially faster than d/n. However, their results also require the axes of $B(c, \boldsymbol{\alpha})$ to decay rapidly (i.e. $a_d/c \to \infty$ quickly) and do not apply to ℓ^2-balls $B(c) = B(c, (1, ..., 1)^T)$ unless $d/n \to 0$. Though these decay conditions are natural for many inverse problems in nonparametric function estimation, they drive the improved minimax rates obtained by Goldenshluger and Tsybakov and may be overly restrictive in other settings, such as the genomics applications discussed in Section 1 above.

2.3. The sequence model

Minimax problems over restricted parameter spaces have been studied extensively in the context of the sequence model. In the sequence model, given an index set J,

$$z_j = \theta_j + \delta_j, \quad j \in J, \tag{5}$$

are observed, $\boldsymbol{\theta} = (\theta_j)_{j \in J}$ is an unknown parameter, and $\boldsymbol{\delta} = (\delta_j)_{j \in J}$ is a random error. The sequence model is extremely flexible, and many existing results about the Gaussian sequence model (where the coordinates of $\boldsymbol{\delta}$ are iid Gaussian random variables) have implications for high-dimensional linear models (Cavalier and Tsybakov, 2002; Pinsker, 1980). However, these results tend to apply in linear models where one conditions on the predictors, as opposed to random predictor models like the one considered here.

In order to prove the main result in this paper (Theorem 1), we study a sequence model with non-Gaussian orthogonally invariant errors that is equivalent to the linear model (1). Goldenshluger and Tsybakov (2001) also studied a non-Gaussian sequence model that derives from a high-dimensional linear model with random predictors, but their results have limitations in settings where $d/n \to \rho > 0$, as discussed in Section 2.2 above. In our analysis, orthogonal invariance is heavily exploited to obtain precise results in any asymptotic setting where $d \to \infty$. This appears to be a key difference between our analysis and Goldenshluger and Tsybakov's.

2.4. Minimax problems over ℓ^2-spheres and orthogonal equivariance

Define the ℓ^2-sphere of radius c, $S(c) = S_d(c) = \{\boldsymbol{\beta} \in \mathbb{R}^d;\ ||\boldsymbol{\beta}|| = c\}$. Though it is common in dense estimation problems to study the minimax risk over ℓ^2-balls $R^{(b)}(c)$, which is one of the primary objects of study here, we find it convenient and informative to consider a closely related quantity, the minimax risk over $S(c)$,

$$R^{(s)}(c) = R^{(s)}_{d,n}(c) = \inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in S(c)} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$$

(the superscript "s" in $R^{(s)}(c)$ stands for "sphere"). For our purposes, the primary significance of considering ℓ^2-spheres comes from connections with orthogonal invariance and equivariance. Let $O(d)$ denote the group of $d \times d$ orthogonal matrices.

Definition 1. An estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X)$ for $\boldsymbol{\beta}$ is *orthogonally equivariant* if

$$U^T \hat{\boldsymbol{\beta}}(\mathbf{y}, X) = \hat{\boldsymbol{\beta}}(\mathbf{y}, XU) \tag{6}$$

for all $U \in O(d)$. □

Orthogonally equivariant estimators are compatible with orthogonal transformations of the predictor basis. They may be appropriate when there is little information carried in the given predictor basis vis-à-vis the outcome; by contrast, knowledge about sparsity is exactly one such piece of information. Indeed, sparsity assumptions generally imply that in the given basis some predictors are significantly more influential than others. Sparse estimators attempt to take advantage of this to improve performance and are typically not orthogonally equivariant.

The concept of equivariance plays an important role in statistical decision theory (e.g. (Berger, 1985), Chapter 6). However, it seems to have received relatively little attention in the context of linear models. Significant aspects of equivariance include: (i) in certain cases, one can show that it suffices to consider equivariant estimators when studying minimax problems and (ii) equivariance may provide a convenient means for identifying minimax estimators. This is basically the content of the Hunt-Stein theorem and both of these features prevail in the present circumstances. To make this more precise, define the class of equivariant estimators

$$\mathscr{E} = \mathscr{E}(n,d) = \{\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\beta}} \text{ is an orthogonally equivariant estimator for } \boldsymbol{\beta}\}$$

and define

$$R^{(e)}(\boldsymbol{\beta}) = R^{(e)}_{d,n}(\boldsymbol{\beta}) = \inf_{\hat{\boldsymbol{\beta}} \in \mathscr{E}} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}).$$

Additionally, let π_c denote the uniform measure on $S(c)$ and let

$$\hat{\boldsymbol{\beta}}_{unif}(c) = \hat{\boldsymbol{\beta}}_{unif}(\mathbf{y}, X; c) = E_{\pi_c}(\boldsymbol{\beta}|\mathbf{y}, X)$$

be the posterior mean of $\boldsymbol{\beta}$ under the assumption that $\boldsymbol{\beta} \sim \pi_c$ is independent of $(\boldsymbol{\epsilon}, X)$. Since, for $U \in O(d)$,

$$U^T \hat{\boldsymbol{\beta}}_{unif}(\mathbf{y}, X; c) = E_{\pi_c}(U^T\boldsymbol{\beta}|\mathbf{y}, X) = E_{\pi_c}(\boldsymbol{\beta}|\mathbf{y}, XU) = \hat{\boldsymbol{\beta}}_{unif}(\mathbf{y}, XU; c),$$

it follows that $\hat{\boldsymbol{\beta}}_{unif}(c) \in \mathscr{E}$. The next result follows directly from the Hunt-Stein theorem and its proof is omitted.

Proposition 1. *Suppose that $||\boldsymbol{\beta}|| = c$. Then*

$$R^{(s)}(c) = R^{(e)}(\boldsymbol{\beta}) = R\{\hat{\boldsymbol{\beta}}_{unif}(c), \boldsymbol{\beta}\}. \tag{7}$$

Furthermore, if $\hat{\boldsymbol{\beta}} \in \mathscr{E}$, then $R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ depends on $\boldsymbol{\beta}$ only through c.

In a sense, Proposition 1 completely solves the minimax problem over $S(c)$. On the other hand, the minimax estimator $\hat{\boldsymbol{\beta}}_{unif}(c)$ is challenging to compute and it is desirable to identify good estimators that have a simpler form. Moreover, though $\hat{\boldsymbol{\beta}}_{unif}(c)$ solves the minimax problem over $S(c)$, it is unclear how $R^{(s)}(c)$ relates to the minimax risk over ℓ^2-balls, which is a more commonly studied quantity in dense estimation problems. Finally, the minimax estimator $\hat{\boldsymbol{\beta}}_{unif}(c)$ depends on $c = ||\boldsymbol{\beta}||$, which is typically unknown in practice. All of these issues must be addressed in order to identify practical estimators that perform well in dense problems for high-dimensional linear models. This is accomplished below, where we show: (i) a linear estimator (ridge regression) is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{unif}(c)$, (ii) $R^{(b)}(c) \sim R^{(s)}(c)$ (i.e. $R^{(b)}(c)/R^{(s)}(c) \to 1$), and (iii) under certain conditions $c = ||\boldsymbol{\beta}||$ may be effectively estimated. Similar results have been obtained for the Gaussian sequence model with iid errors (Beran, 1996; Marchand, 1993). Our results rely on an inequality of Marchand's (Proposition 11 below) and extend Marchand's and Beran's results to linear models with Gaussian predictors.

Proposition 1 and the related discussion imply that equivariant estimators have certain nice properties and are closely linked with dense estimation problems. On the other hand, the next result describes some of the limitations of orthogonally equivariant estimators when $d > n$ and is indicative of some of the challenges inherent in dense estimation problems beyond the consistency requirement $d/n \to 0$.

Lemma 1. *Suppose that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}, X) \in \mathscr{E}$. Then $\hat{\boldsymbol{\beta}}$ is orthogonal to the null-space of X.*

Proof. Suppose that $\operatorname{rank}(X) = r < d$ and let $X = UDV^T$ be the singular value decomposition of X, where $U \in O(n)$, $V \in O(d)$,

$$D = \left(\begin{array}{cc} D_0 & 0 \\ 0 & 0 \end{array} \right)$$

is an $n \times d$ matrix, and D_0 is an $r \times r$ diagonal matrix with rank r. Let V_0 denote the first r columns of V and let V_1 denote the remaining $d - r$ columns of V. Finally, suppose that $W_1 \in O(d-r)$ and let

$$W = \left(\begin{array}{cc} I & 0 \\ 0 & W_1 \end{array} \right) \in O(d).$$

Then the null space of X is equal to the column space of V_1 and it suffices to show that $V_1^T\hat{\boldsymbol{\beta}} = 0$. By equivariance,

$$\hat{\boldsymbol{\beta}} = VW\hat{\boldsymbol{\beta}}(\mathbf{y}, XVW) = VW\hat{\boldsymbol{\beta}}(\mathbf{y}, UD). \tag{8}$$

Thus,

$$V_1^T\hat{\boldsymbol{\beta}} = V_1^TVW\hat{\boldsymbol{\beta}}(\mathbf{y}, UD) = (0 \;\; W)\,\hat{\boldsymbol{\beta}}(\mathbf{y}, UD). \tag{9}$$

Since $\hat{\boldsymbol{\beta}}(\mathbf{y}, UD)$ does not depend on W and (9) holds for all $W \in O(d-r)$, it follows that $V_1^T\hat{\boldsymbol{\beta}} = 0$, as was to be shown. □

Lemma 1 is a non-estimability result for orthogonally equivariant estimators. It will be used in Sections 3.3 and 6 below.

2.5. Linear estimators: Ridge regression

Linear estimators play an important role in dense estimation problems in many statistical settings. Fundamental references include (James and Stein, 1961; Pinsker, 1980; Stein, 1955). Pinsker (1980) showed that under certain conditions, linear estimators in the Gaussian sequence model are asymptotically minimax over ℓ^2-ellipsoids. In the linear model, linear estimators have the form $\hat{\boldsymbol{\beta}} = A\mathbf{y}$, where A is a data-dependent $d \times n$ matrix, and they are convenient because of their simplicity. Define the ridge regression estimator

$$\hat{\boldsymbol{\beta}}_r(c) = (X^TX + d/c^2I)^{-1}X^T\mathbf{y}, \quad c \in [0, \infty].$$

By convention, we take $\hat{\boldsymbol{\beta}}_r(0) = 0$ and $\hat{\boldsymbol{\beta}}_r(\infty) = \hat{\boldsymbol{\beta}}_{ols} = (X^TX)^{-1}X^T\mathbf{y}$ to be the ordinary least squares (OLS) estimator. Furthermore, throughout the paper, if a matrix A is not invertible, then A^{-1} is taken to be its Moore-Penrose pseudoinverse (thus, the OLS estimator is defined for all d, n). Clearly, $\hat{\boldsymbol{\beta}}_r(c)$ is a linear estimator. Furthermore, it is easy to check that $\hat{\boldsymbol{\beta}}_r(c) \in \mathscr{E}$.

Dicker (2012) studied finite sample and asymptotic properties of $R\{\hat{\boldsymbol{\beta}}_r(c), \boldsymbol{\beta}\}$. Some of these properties will be used in this paper and are summarized presently.

2.5.1. *Oracle estimators*

Define the oracle ridge regression estimator

$$\hat{\boldsymbol{\beta}}_r^* = \hat{\boldsymbol{\beta}}_r(||\boldsymbol{\beta}||).$$

This estimator is called an oracle estimator because it depends on $||\boldsymbol{\beta}||$, which is typically unknown. Proposition 5 of (Dicker, 2012) implies

$$R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) = \inf_{c \in [0,\infty]} R\{\hat{\boldsymbol{\beta}}_r(c), \boldsymbol{\beta}\} = E\mathrm{tr}(X^TX + d/||\boldsymbol{\beta}||^2 I)^{-1} \tag{10}$$

and, furthermore,

$$R\{\hat{\boldsymbol{\beta}}_r(||\boldsymbol{\beta}||), \boldsymbol{\beta}_0\} \leq R\{\hat{\boldsymbol{\beta}}_r(||\boldsymbol{\beta}||), \boldsymbol{\beta}\}, \text{ if } ||\boldsymbol{\beta}_0|| \leq ||\boldsymbol{\beta}||. \tag{11}$$

The next result gives an expression for the asymptotic predictive risk of $\hat{\boldsymbol{\beta}}_r^*$. Its proof relies heavily on properties of the Marčenko-Pastur distribution (Bai, 1993; Marčenko and Pastur, 1967).

Proposition 2 (Proposition 8 from (Dicker, 2012))**.** *Suppose that $0 < \rho^- \leq d/n \leq \rho^+ < \infty$ for some fixed constants $\rho^-, \rho^+ \in \mathbb{R}$ and define*

$$r_{>0}(\rho, c) = \frac{1}{2\rho}\left[c^2(\rho - 1) - \rho + \sqrt{\{c^2(\rho-1) - \rho\}^2 + 4c^2\rho^2}\right].$$

(a) If $0 < \rho^- < \rho^+ < 1$ or $1 < \rho^- < \rho^+ < \infty$ and $n - d > 5$, then

$$\left| R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) - r_{>0}(d/n, ||\boldsymbol{\beta}||) \right| = O\left(\frac{||\boldsymbol{\beta}||^2}{||\boldsymbol{\beta}||^2 + 1} n^{-1/4} \right).$$

(b) If $0 < \rho^- < 1 < \rho^+ < \infty$, then

$$\left| R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) - r_{>0}(d/n, ||\boldsymbol{\beta}||) \right| = O(||\boldsymbol{\beta}||^2 n^{-5/48}).$$

Notice that Proposition 2 implies the asymptotic predictive risk of $\hat{\boldsymbol{\beta}}_r^*$ is non-vanishing if $d/n \to \rho > 0$. The main results in this paper are essentially asymptotic optimality results for $\hat{\boldsymbol{\beta}}_r^*$. In particular, we show that $\hat{\boldsymbol{\beta}}_r^*$ is asymptotically minimax over ℓ^2-balls and ℓ^2-spheres, and asymptotically optimal among the class of orthogonally equivariant estimators. Combined with Propositions 2-3, these results immediately yield sharp asymptotic for $R^{(b)}(c)$, $R^{(s)}(c)$, and $R^{(e)}(\boldsymbol{\beta})$.

Taking a Bayesian point-of-view, our optimality results for $\hat{\boldsymbol{\beta}}_r^*$ are not surprising. Indeed, in Section 2.3 we observed that if $||\boldsymbol{\beta}|| = c$, then $\hat{\boldsymbol{\beta}}_{unif}(c) = E_{\pi_c}(\boldsymbol{\beta}|\mathbf{y}, X)$ is minimax over $S(c)$ and is optimal among orthogonally equivariant estimators for $\boldsymbol{\beta}$. On the other hand, if $||\boldsymbol{\beta}|| = c$, then the oracle ridge estimator $\hat{\boldsymbol{\beta}}_r^* = \hat{\boldsymbol{\beta}}_r(c) = E_{N(0,c^2/dI)}(\boldsymbol{\beta}|\mathbf{y}, X)$ may be interpreted as the posterior mean of $\boldsymbol{\beta}$ under the assumption that $\boldsymbol{\beta} \sim N\{0, (c^2/d)I\}$ is independent of $\boldsymbol{\epsilon}$ and X. Furthermore, if d is large, then the normal distribution $N\{0, (c^2/d)I\}$ is "close" to π_c (there is an enormous body of literature that makes this idea more precise – Diaconis and Freedman (1987) attribute early work to Borel (1914) and Lévy (1922)). Thus, it is reasonable to expect that $\hat{\boldsymbol{\beta}}_{unif}(c) \approx \hat{\boldsymbol{\beta}}_r(c)$ and that, asymptotically, the oracle ridge estimator shares the optimality properties of $\hat{\boldsymbol{\beta}}_{unif}(c)$, which is indeed the case.

2.5.2. *Adaptive estimators*

Adaptive ridge estimators will also be discussed in this paper. As mentioned above, $||\boldsymbol{\beta}||$ is typically unknown; hence, $\hat{\boldsymbol{\beta}}_r^*$ is typically non-implementable. However, $\hat{\boldsymbol{\beta}}_r^*$ may be approximated by an adaptive estimator where $||\boldsymbol{\beta}||$ is replaced with an estimate – this estimator "adapts" to the unknown quantity $||\boldsymbol{\beta}||$. Define

$$\widehat{||\boldsymbol{\beta}||}^2 = \max\left\{\frac{||\mathbf{y}||^2}{n} - 1, 0\right\}$$

and define the adaptive ridge estimator

$$\check{\boldsymbol{\beta}}_r^* = \hat{\boldsymbol{\beta}}_r(\widehat{||\boldsymbol{\beta}||}). \tag{12}$$

Note that $\widehat{||\boldsymbol{\beta}||}^2$ is a consistent estimator of $||\boldsymbol{\beta}||^2$, as $n \to \infty$.

Proposition 3. *Suppose that $0 < \rho^- \leq d/n \leq \rho^+ < 1$ for some fixed constants $\rho^-, \rho^+ \in \mathbb{R}$. If $n - d > 5$, then*

$$\left|R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) - R(\check{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta})\right| = O\left(\frac{1}{||\boldsymbol{\beta}||^2 + 1} n^{-1/2}\right).$$

The proof of Proposition 3 is nearly identical to the proof of Proposition 10 from (Dicker, 2012) and is omitted. Proposition 3 implies that if $d/n \to \rho \in (0,1)$, then the adaptive ridge estimator has nearly the same asymptotic risk as the oracle ridge estimator. Note the restriction $d/n < 1$ in Proposition 3. This restriction also appears in (Dicker, 2012), where $\mathrm{Var}(\epsilon_i) = \sigma^2$ is unknown and the signal-to-noise ratio $||\boldsymbol{\beta}||^2/\sigma^2$, as opposed to $||\boldsymbol{\beta}||^2$, is the quantity that must be estimated to obtain an adaptive ridge estimator; in this context, $d/n < 1$ is a fairly natural condition for estimating σ^2. It is possible to extend Proposition 3 to settings where $d/n > 1$. However, if $d/n > 1$, then the corresponding error term in Proposition 3 is no longer uniformly bounded in $||\boldsymbol{\beta}||^2$. Additionally, notice that Proposition 3 does not apply to settings where $d/n \to 0$. A more careful analysis may lead to extensions in this direction as well. Since adaptive estimation is not the main focus of this article, these issues are not pursued further here; however, future research into these issues may prove interesting.

2.6. Outline of the paper

The main results of the paper are stated in Section 3. Most of these results follow from Theorem 1, which is stated at the beginning of the section. The remainder of the paper is devoted to proving Theorem 1. In Section 4, the equivalence between the linear model and the sequence model is formalized. The first part of Theorem 1, which applies to the setting where $d \leq n$, is proved in Section 5. This part of the proof involves converting error bounds for the Gaussian sequence model with iid errors into useful bounds for the relevant non-Gaussian sequence model. The second part of Theorem 1 ($d > n$) is proved in Section 6. When $d > n$, X^TX does not have full rank. The major steps in the proof for $d > n$ involve reducing the problem to a full rank problem.

3. Main results

The results in this section are presented in terms of the linear model. However, most have equivalent formulations in terms of the sequence model introduced in Section 4 below.

Theorem 1. *Suppose that $n > 2$ and let $s_1 \geq \cdots \geq s_{d\wedge n} > 0$ denote the nonzero (with probability 1) eigenvalues of $(X^TX)^{-1}$.*

(a) If $d \leq n$, then

$$\left|R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) - R^{(e)}(\boldsymbol{\beta})\right| \leq \frac{1}{d} E\left\{\frac{s_1}{s_d} \mathrm{tr}\left(X^TX + \frac{d}{||\boldsymbol{\beta}||^2} I\right)^{-1}\right\}$$

(b) If $d > n$, then

$$\begin{aligned} \left| R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) - R^{(e)}(\boldsymbol{\beta}) \right| &\leq \frac{1}{n} E\left\{ \frac{s_1}{s_n} \mathrm{tr}\left(XX^T + \frac{d}{||\boldsymbol{\beta}||^2} I \right)^{-1} \right\} \\ &\quad + 2\frac{d-n}{n-2} \frac{1}{||\boldsymbol{\beta}||^2} E\mathrm{tr}\left(XX^T + \frac{d}{||\boldsymbol{\beta}||^2} I \right)^{-2}. \end{aligned}$$

From (10) and Proposition 1, it is clear that $R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta})$ and $R^{(e)}(\boldsymbol{\beta})$ are finite. Moreover, basic properties of the Wishart and inverse Wishart distributions imply that the upper bounds in Theorem 1 are finite, provided $|n-d| > 1$; when $|n-d| \leq 1$, these bounds are infinite. However, if $|n-d| \leq 1$, then the inequalities $R_{d,n}(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) \leq R_{d,n-1}(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta})$ and $R_{d,n}^{(e)}(\boldsymbol{\beta}) \leq R_{d,n-1}^{(e)}(\boldsymbol{\beta})$ may be combined with Theorem 1 (b) to obtain nontrivial bounds.

In what remains of this section, we discuss some of the consequences of Theorem 1 and related results in three asymptotic settings: $d/n \to 0$ (with $d \to \infty$, as well), $d/n \to \rho \in (0, \infty)$, and $d/n \to \infty$.

3.1. $d/n \to 0$

Proposition 4. *Define*

$$r_0(\rho, c) = \frac{c^2 \rho}{c^2 + \rho}.$$

If $d/n \to 0$ and $d \to \infty$, then

$$R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) \sim R^{(e)}(\boldsymbol{\beta}) \sim R^{(s)}(||\boldsymbol{\beta}||) \sim R^{(b)}(||\boldsymbol{\beta}||) \sim r_0(d/n, ||\boldsymbol{\beta}||)$$

uniformly for $\boldsymbol{\beta} \in \mathbb{R}^d$.

Proof. If $d + 1 < n$, then (10) and Jensen's inequality imply that

$$\frac{d/n}{1 + d/(n||\boldsymbol{\beta}||^2)} \leq R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) \leq \frac{d/n}{1 - (d+1)/n + d/(n||\boldsymbol{\beta}||^2)}. \tag{13}$$

It follows that $R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) \sim r_0(d/n, ||\boldsymbol{\beta}||)$. By Theorem 1, in order to prove

$$R^{(e)}(\boldsymbol{\beta}) \sim r_0(d/n, ||\boldsymbol{\beta}||), \tag{14}$$

it suffices to show that

$$\frac{1}{d} E\left\{ \frac{s_1}{s_d} \mathrm{tr}\left(X^T X + d/||\boldsymbol{\beta}||^2 I \right)^{-1} \right\} = o\{r_0(d/n, ||\boldsymbol{\beta}||)\}.$$

But this is clear:

$$\begin{aligned}\frac{1}{d}E\left\{\frac{s_1}{s_d}\mathrm{tr}\left(X^TX+\frac{d}{||\boldsymbol{\beta}||^2}I\right)^{-1}\right\} &\leq \frac{||\boldsymbol{\beta}||^2}{d(||\boldsymbol{\beta}||^2+d/n)} \\ &\quad \cdot E\left\{\frac{s_1}{s_d}\left(ds_1+\frac{d}{n}\right)\right\} \\ &= O\left\{\frac{1}{d}r_0(d/n,||\boldsymbol{\beta}||)\right\} \\ &= o\{r_0(d/n,||\boldsymbol{\beta}||)\},\end{aligned} \tag{15}$$

where we have used the facts $E(s_1^k) = O(n^{-k})$ and $E(s_d^{-k}) = O(n^k)$ (Lemma A2, (Dicker, 2012)). Thus, (14). Since $R^{(s)}(||\boldsymbol{\beta}||) = R^{(e)}(\boldsymbol{\beta})$, all that is left is to prove is $R^{(b)}(||\boldsymbol{\beta}||) \sim R^{(s)}(||\boldsymbol{\beta}||)$. This follows because

$$R^{(s)}(||\boldsymbol{\beta}||) \leq R^{(b)}(||\boldsymbol{\beta}||) \leq R(\hat{\boldsymbol{\beta}}_r^*,\boldsymbol{\beta}) \sim R^{(s)}(||\boldsymbol{\beta}||), \tag{16}$$

where we have used (11) to obtain the second inequality. □

The asymptotic risk $r_0(\rho, c)$ appears frequently in the analysis of linear estimators for the Gaussian sequence model (Pinsker, 1980) and is often referred to as the "linear minimax risk." The condition $d \to \infty$ in Proposition 4 is important because it drives the approximation $\pi_c \approx N(0, c^2/dI)$, which enables us to conclude $R^{(e)}(\boldsymbol{\beta}) \sim R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta})$ (re: the discussion at the end of Section 2.4). Notice that $\lim_{d/n\to 0} r_0(\rho, c) = 0$. Thus, the minimax risk vanishes when $d/n \to 0$.

Proposition 4 implies that the ridge estimator $\hat{\boldsymbol{\beta}}_r^*$ is asymptotically minimax if $d/n \to 0$ and $d \to \infty$. On the other hand, other simple linear estimators are also asymptotically minimax in this setting. Define the estimator

$$\hat{\boldsymbol{\beta}}_{scal}^* = \frac{1-(d+1)/n}{1-(d+1)/n+d/(n||\boldsymbol{\beta}||^2)}\hat{\boldsymbol{\beta}}_{ols}.$$

Note that $\hat{\boldsymbol{\beta}}_{scal}^*$ is a scalar multiple of the OLS estimator and that $\hat{\boldsymbol{\beta}}_{scal}^*$ is defined for all d, n since $\hat{\boldsymbol{\beta}}_{ols}$ is defined using pseudoinverses. Various versions of $\hat{\boldsymbol{\beta}}_{scal}$ have been studied previously (Baranchik, 1973; Brown, 1990; Stein, 1960). Dicker (2012) showed that if $d+1 < n$, then

$$\begin{aligned}R(\hat{\boldsymbol{\beta}}_r^*,\boldsymbol{\beta}) &\leq R(\hat{\boldsymbol{\beta}}_{scal}^*,\boldsymbol{\beta}) = \frac{d/n}{1-(d+1)/n+d/(n||\boldsymbol{\beta}||^2)} \\ &\leq R(\hat{\boldsymbol{\beta}}_{ols},\boldsymbol{\beta}) = \frac{d/n}{1-(d+1)/n}.\end{aligned} \tag{17}$$

The following corollary to Proposition 4 follows immediately.

Corollary 1. *(a) If $d/n \to 0$ and $d \to \infty$, then*

$$R(\hat{\boldsymbol{\beta}}^*_{scal}, \boldsymbol{\beta}) \sim R^{(b)}(||\boldsymbol{\beta}||)$$

uniformly for $\boldsymbol{\beta} \in \mathbb{R}^d$.

(b) If $d/n \to 0$, $d \to \infty$, and $d/(n||\boldsymbol{\beta}||^2) \to s \geq 0$, then

$$\frac{R(\hat{\boldsymbol{\beta}}_{ols}, \boldsymbol{\beta})}{R^{(b)}(||\boldsymbol{\beta}||)} \to 1 + s.$$

In other words, if $d/n \to 0$ and $d \to \infty$, then $\hat{\boldsymbol{\beta}}_{scal}$ is asymptotically minimax over ℓ^2-balls (and, moreover, asymptotically equivalent to $\hat{\boldsymbol{\beta}}^*_r$). Furthermore, the OLS estimator may be asymptotically minimax over ℓ^2-balls, but this depends on the magnitude of the signal $\boldsymbol{\beta}$: If $||\boldsymbol{\beta}||^2$ is large, then the OLS estimator is asymptotically minimax; if $||\boldsymbol{\beta}||^2$ is small, then it is not.

3.2. $d/n \to \rho \in (0, \infty)$

The setting where $d/n \to \rho \in (0, \infty)$ may be the most interesting one for the dense estimation problems considered here. The minimax risk is non-vanishing in this setting; however, informative closed form expressions for the limiting minimax risk are available. Moreover, differences between the linear estimators $\hat{\boldsymbol{\beta}}^*_{scal}$ and $\hat{\boldsymbol{\beta}}^*_r$ which are insignificant when $d/n \to 0$ become pronounced when $d/n \to \rho \in (0, \infty)$. These differences are largely attributable to the spectral distribution of $n^{-1}X^TX$, which is asymptotically trivial if $d/n \to 0$ and converges to the Marčenko-Pastur distribution (Marčenko and Pastur, 1967) if $d/n \to \rho \in (0, \infty)$.

Proposition 5. *Suppose that $\rho \in (0, \infty)$ and let $R^*(\boldsymbol{\beta})$ denote any of $R(\hat{\boldsymbol{\beta}}^*_r, \boldsymbol{\beta})$, $R^{(e)}(\boldsymbol{\beta})$, $R^{(s)}(||\boldsymbol{\beta}||)$, or $R^{(b)}(\boldsymbol{\beta})$. If $\rho \neq 1$, then*

$$\lim_{d/n \to \rho} \sup_{\boldsymbol{\beta} \in \mathbb{R}^d} |R^*(\boldsymbol{\beta}) - r_{>0}(d/n, ||\boldsymbol{\beta}||)| = 0, \tag{18}$$

where $r_{>0}(\rho, c)$ is defined in Proposition 2 above. Furthermore, as $d/n \to \rho$,

$$R(\hat{\boldsymbol{\beta}}^*_r, \boldsymbol{\beta}) \sim R^{(e)}(\boldsymbol{\beta}) \sim R^{(s)}(||\boldsymbol{\beta}||) \sim R^{(b)}(||\boldsymbol{\beta}||) \sim r_{>0}(d/n, ||\boldsymbol{\beta}||). \tag{19}$$

If $\rho \neq 1$, then the implied convergence in (19) holds uniformly for $\boldsymbol{\beta} \in \mathbb{R}^d$; if $\rho = 1$, then the convergence is uniform over $B(c)$ for any fixed $c \in (0, \infty)$.

Proof. Proposition 2 implies that $|R(\hat{\boldsymbol{\beta}}^*_r, \boldsymbol{\beta}) - r_{>0}(d/n, ||\boldsymbol{\beta}||)| \to 0$ and $R(\hat{\boldsymbol{\beta}}^*_r, \boldsymbol{\beta}) \sim r_{>0}(d/n, ||\boldsymbol{\beta}||)$, with the appropriate uniformity conditions when $\rho \neq 1$ or $\rho = 1$. For $\rho \leq 1$, the asymptotic equivalences $|R^{(e)}(\boldsymbol{\beta}) - R(\hat{\boldsymbol{\beta}}^*_r, \boldsymbol{\beta})| \to 0$ and $R^{(e)}(\boldsymbol{\beta}) \sim R(\hat{\boldsymbol{\beta}}^*_r, \boldsymbol{\beta})$ follow from (13) and (15); to prove the equivalences for $\rho > 1$, notice that

$$\begin{array}{l} \frac{1}{n}E\left\{\frac{s_1}{s_n}\mathrm{tr}(XX^T + d/c^2I)^{-1}\right\} \\ \qquad +2\frac{d-n}{c^2(n-2)}E\mathrm{tr}(XX^T + d/c^2I)^{-2} \end{array} = O\left\{\frac{||\boldsymbol{\beta}||^2}{n(||\boldsymbol{\beta}||^2+1)}\right\}.$$

Since $R^{(e)}(\boldsymbol{\beta}) = R^{(s)}(||\boldsymbol{\beta}||)$, it suffices to show that

$$\lim_{d/n\to\rho} \sup_{\boldsymbol{\beta}\in\mathbb{R}^d} \left|R^{(s)}(||\boldsymbol{\beta}||) - R^{(b)}(||\boldsymbol{\beta}||)\right| = 0$$

and that $R^{(s)}(||\boldsymbol{\beta}||) \sim R^{(b)}(||\boldsymbol{\beta}||)$ uniformly for $\boldsymbol{\beta} \in \mathbb{R}^d$ in order to prove the proposition; both follow from (16). □

Two types of asymptotic equivalence are addressed in Proposition 5: differences (18) and quotients (19). The equivalence (18) is more informative for large $||\boldsymbol{\beta}||$; (19) is more informative for small $||\boldsymbol{\beta}||$. Notice that for fixed $||\boldsymbol{\beta}|| = c \in (0, \infty)$, $\lim_{d/n\to\rho} r_{>0}(d/n, c) = r_{>0}(\rho, c) > 0$ and it follows that (18) and (19) are equivalent.

For $d/n \to 0$, we saw that $\hat{\boldsymbol{\beta}}^*_{scal}$ and $\hat{\boldsymbol{\beta}}^*_r$ were asymptotically equivalent (and that, in some instance, both were also asymptotically equivalent to the OLS estimator; Corollary 1). When $d/n \to \rho \in (0, \infty)$, $\hat{\boldsymbol{\beta}}^*_r$ and $\hat{\boldsymbol{\beta}}^*_{scal}$ are not asymptotically equivalent. Indeed, (17) implies that for $d/n \to 0$, we have

$$R(\hat{\boldsymbol{\beta}}^*_{scal}, \boldsymbol{\beta}) \sim r_{scal}(d/n, ||\boldsymbol{\beta}||),$$

where

$$r_{scal}(\rho, c) = \frac{1-\rho}{1-\rho+\rho/c^2}.$$

One easily checks that for $\rho > 0$, $r_{>0}(\rho, c) \leq r_{scal}(\rho, c)$ with equality if and only if $c = 0$. Thus, $\hat{\boldsymbol{\beta}}^*_{scal}$ is not asymptotically minimax over ℓ^2-balls when $d/n \to \rho \in (0, \infty)$.

Despite its suboptimal performance, the estimator $\hat{\boldsymbol{\beta}}^*_{scal}$ may be useful in certain situations. Indeed, if $\mathrm{Cov}(\mathbf{x}_i) = \Sigma \neq I$, then it is straightforward to implement a modified version of $\hat{\boldsymbol{\beta}}^*_{scal}$ with similar properties (replace $||\boldsymbol{\beta}||^2$ in $\hat{\boldsymbol{\beta}}^*_{scal}$ with $\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}$); on the other hand, if Σ is unknown and a norm-consistent estimator for Σ is not available, then this may have a more dramatic effect on the ridge estimator $\hat{\boldsymbol{\beta}}^*_r$. This is discussed in detail in (Dicker, 2012), where it is argued that in dense problems where little is known about $\mathrm{Cov}(\mathbf{x}_i)$, an appropriately modified version of $\hat{\boldsymbol{\beta}}^*_{scal}$ is a reasonable alternative to ridge regression (note, for instance, that $R(\hat{\boldsymbol{\beta}}^*_{scal}, \boldsymbol{\beta})/R(\hat{\boldsymbol{\beta}}^*_r, \boldsymbol{\beta}) = O(1)$ if $d/n \to \rho \in (0, \infty)$).

3.3. $d/n \to \infty$

Theorem 1 plays a crucial role in our asymptotic analysis when $d/n \to \rho < \infty$. It is less relevant in the setting where $d/n \to \infty$. Instead, Lemma 1 from Section 2.4 plays the key role. We have the following proposition.

Proposition 6. *Suppose that $d > n$ and that $\hat{\boldsymbol{\beta}} \in \mathscr{E}$. Then*

$$R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq \frac{d-n}{d}||\boldsymbol{\beta}||^2.$$

Proof. Let $X = UDV^T$ be the singular value decomposition of X, as in the proof of Lemma 1. Let V_0 and V_1 be the first r and the remaining $d-r$ columns of V, respectively, where $r = \text{rank}(X)$ (note that $r = n$ with probability 1). Then

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= E||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 \\ &= E||V_0^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2 + E||V_1^T\boldsymbol{\beta}||^2 \qquad (20) \\ &\geq E||V_1^T\boldsymbol{\beta}||^2 \\ &= \frac{d-n}{n}||\boldsymbol{\beta}||^2, \qquad (21) \end{aligned}$$

where (20) follows from Lemma 1 and (21) follows from symmetry. □

The proof of Proposition 6 essentially implies that for $d > n$, the squared bias of an equivariant estimator must be at least $||\boldsymbol{\beta}||^2(d-n)/d$. This highlights one of the major challenges in high-dimensional dense estimation problems, especially in settings where $d \gg n$. The next proposition, which is the main result in this subsection, implies that if $d/n \to \infty$, then the trivial estimator $\hat{\boldsymbol{\beta}}_{null} = 0$ is asymptotically minimax. In a sense, this means that in dense problems $\boldsymbol{\beta}$ is completely non-estimable when $d/n \to \infty$.

Proposition 7. *Let $\hat{\boldsymbol{\beta}}_{null} = 0$. Then $R(\hat{\boldsymbol{\beta}}_{null}, \boldsymbol{\beta}) = ||\boldsymbol{\beta}||^2$. Furthermore, if $d/n \to \infty$, then*

$$R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) \sim R^{(e)}(\boldsymbol{\beta}) \sim R^{(s)}(||\boldsymbol{\beta}||) \sim R^{(b)}(||\boldsymbol{\beta}||) \sim R(\hat{\boldsymbol{\beta}}_{null}, \boldsymbol{\beta}) \sim ||\boldsymbol{\beta}||^2$$

uniformly for $\boldsymbol{\beta} \in \mathbb{R}^d$.

Proof. Clearly, $R(\hat{\boldsymbol{\beta}}_{null}, \boldsymbol{\beta}) = ||\boldsymbol{\beta}||^2$. It follows from Proposition 6 that for $d > n$,

$$\begin{aligned} \frac{d-n}{n}||\boldsymbol{\beta}||^2 \leq R^{(e)}(\boldsymbol{\beta}) = R^{(s)}(||\boldsymbol{\beta}||^2) &\leq R^{(b)}(||\boldsymbol{\beta}||^2) \\ &\leq R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) \leq R(\hat{\boldsymbol{\beta}}_{null}, \boldsymbol{\beta}) = ||\boldsymbol{\beta}||^2. \end{aligned}$$

The proposition follows by dividing by $||\boldsymbol{\beta}||^2$ and taking $d/n \to \infty$. □

3.4. Adaptive estimators

The results in Section 3.1-3.3 imply that the oracle ridge estimator $\hat{\boldsymbol{\beta}}_r^* = \hat{\boldsymbol{\beta}}_r(||\boldsymbol{\beta}||)$ is asymptotically minimax over ℓ^2-balls and ℓ^2-spheres and is asymptotically optimal among equivariant estimators for $\boldsymbol{\beta}$ in any asymptotic setting where $d \to \infty$. The next result describes asymptotic optimality properties of the adaptive ridge estimator $\check{\boldsymbol{\beta}}_r^*$ (defined in (12)), which does not depend on $||\boldsymbol{\beta}||$.

Proposition 8. *Suppose that $\rho \in (0,1)$ and let $R^*(\boldsymbol{\beta})$ denote any of $R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta})$, $R^{(e)}(\boldsymbol{\beta})$, $R^{(s)}(||\boldsymbol{\beta}||)$, $R^{(b)}(\boldsymbol{\beta})$, or $r_{>0}(d/n, ||\boldsymbol{\beta}||)$. Let $\{a_n\}_{n=1}^{\infty} \subseteq \mathbb{R}$ denote a sequence of positive real numbers such that $a_n n^{1/2} \to \infty$. Then*

$$\lim_{d/n\to\rho} \sup_{\boldsymbol{\beta}\in\mathbb{R}^d} \left| R^*(\boldsymbol{\beta}) - R(\check{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) \right| = 0 \text{ and } \lim_{d/n\to\rho} \sup_{||\boldsymbol{\beta}||^2 \geq a_n} \frac{R(\check{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta})}{R^*(\boldsymbol{\beta})} = 1.$$

Proposition 8 follows immediately from Propositions 3 and 5. The restriction $||\boldsymbol{\beta}||^2 \gg n^{1/2}$ in the second part of Proposition 8 is related to the fact that for $d/n \to \rho \in (0,\infty)$, $R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) = O(||\boldsymbol{\beta}||^2)$ and the error bound in Proposition 3 is $O(n^{-1/2})$. As discussed in Section 2.5.2, more detailed results on adaptive ridge estimators are likely possible (that may apply, for instance, in settings where $d/n \to 0$ or $d/n \to \rho \geq 1$), but this not pursued further here.

4. An equivalent sequence model

The rest of the paper is devoted to proving Theorem 1. In this section and Section 5, we assume that $d \leq n$. In Section 6, we address the case where $d > n$. The major goal in this section is to relate the linear model (1) to an equivalent non-Gaussian sequence model.

4.1. The model

Let Σ be a random orthogonally invariant $m \times m$ positive definite matrix with rank m, almost surely (by orthogonally invariant, we mean that Σ and $U\Sigma U^T$ have the same distribution for any $U \in O(m)$). Additionally, let $\boldsymbol{\delta}_0 \sim N(0, I_m)$ be a d-dimensional Gaussian random vector that is independent of Σ. Recall that in the sequence model (5), the vector $\mathbf{z} = (z_j)_{j\in J} = \boldsymbol{\theta} + \boldsymbol{\delta}$ is observed and J is an index set. In the formulation considered here, $J = \{1, ..., m\}$, $\boldsymbol{\delta} = \Sigma^{1/2}\boldsymbol{\delta}_0$, and Σ is observed along with $\mathbf{z}$. Thus, the available data are $(\mathbf{z}, \Sigma)$ and

$$\mathbf{z} = \boldsymbol{\theta} + \boldsymbol{\delta} = \boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}_0 \in \mathbb{R}^m. \tag{22}$$

Notice that $\boldsymbol{\delta}$ is in general non-Gaussian. However, conditional on Σ, $\boldsymbol{\delta}$ is a Gaussian random vector with covariance Σ. We are interested in the risk for estimating $\boldsymbol{\theta}$ under squared error loss. For an estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{z}, \Sigma)$, this is defined by

$$\tilde{R}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}||\hat{\boldsymbol{\theta}}(\mathbf{z}, \Sigma) - \boldsymbol{\theta}||^2 = E_{\boldsymbol{\theta}}||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2,$$

where the expectation is taken with respect to $\boldsymbol{\delta}_0$ and Σ (we use "$\sim$," as in $\tilde{R}$, to denote quantities related to the sequence model, as opposed to the linear model).

4.2. Equivariance and optimality concepts

Most of the key concepts initially introduced in the context of the linear model have analogues in the sequence model (22). In this subsection, we describe some that will be used in our proof of Theorem 1.

Definition 2. Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{z}, \Sigma)$ be an estimator for $\boldsymbol{\theta}$. Then $\hat{\boldsymbol{\theta}}$ is an *orthogonally equivariant* estimator for $\boldsymbol{\theta}$ if

$$U\hat{\boldsymbol{\theta}}(\mathbf{z}, \Sigma) = \hat{\boldsymbol{\theta}}(U\mathbf{z}, U^T\Sigma U)$$

for all $U \in O(d)$. □

Let

$$\tilde{\mathscr{E}} = \tilde{\mathscr{E}}_d = \{\hat{\boldsymbol{\theta}};\ \hat{\boldsymbol{\theta}} \text{ is an orthogonally equivariant estimator for } \boldsymbol{\theta}\}$$

denote the class of orthogonally equivariant estimators for $\boldsymbol{\theta}$. Also define the posterior mean for $\boldsymbol{\theta}$ under the assumption that $\boldsymbol{\theta} \sim \pi_c$,

$$\hat{\boldsymbol{\theta}}_{unif}(c) = E_{\pi_c}(\boldsymbol{\theta}|\mathbf{z}, \Sigma)$$

and the posterior mean under the assumption that $\boldsymbol{\theta} \sim N(0, c^2/mI)$.

$$\hat{\boldsymbol{\theta}}_r(c) = E_{N(0,c^2/mI)}(\boldsymbol{\theta}|\mathbf{z}, \Sigma) = c^2/d\left\{\Sigma + c^2/mI\right\}^{-1}\mathbf{z}$$

(for both of these Bayes estimators we assume that $\boldsymbol{\theta}$ is independent of $\boldsymbol{\delta}_0$ and Σ). The estimators $\hat{\boldsymbol{\theta}}_{unif}(c)$ and $\hat{\boldsymbol{\theta}}_r(c)$ for $\boldsymbol{\theta}$ are analogous to the estimators $\hat{\boldsymbol{\beta}}_{unif}(c)$ and $\hat{\boldsymbol{\beta}}_r(c)$ for $\boldsymbol{\beta}$, respectively. Moreover, they are both orthogonally equivariant, i.e. $\hat{\boldsymbol{\theta}}_{unif}(c), \hat{\boldsymbol{\theta}}_r(c) \in \tilde{\mathscr{E}}$, and $\hat{\boldsymbol{\theta}}_r(c)$ is a linear estimator. Now define the minimal equivariant risk for the sequence model

$$\tilde{R}^{(e)}(\boldsymbol{\theta}) = \tilde{R}^{(e)}_m(\boldsymbol{\theta}) = \inf_{\hat{\boldsymbol{\theta}} \in \mathscr{E}_{seq}} \tilde{R}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$$

and the minimax risk over the ℓ^2-sphere of radius c,

$$\tilde{R}^{(s)}(c) = \tilde{R}_m^{(s)}(c) = \inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in S(c)} \tilde{R}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}),$$

where the infimum above is taken over all measurable estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{z}, \Sigma)$. The Hunt-Stein theorem yields the following result, which is entirely analogous to Proposition 1.

Proposition 9. *Suppose that $||\boldsymbol{\theta}|| = c$. Then*

$$\tilde{R}^{(s)}(c) = \tilde{R}^{(e)}(\boldsymbol{\theta}) = \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\}.$$

Furthermore, if $\hat{\boldsymbol{\theta}} \in \tilde{\mathscr{E}}$, then $\tilde{R}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ only through c.

4.3. Equivalence of the sequence model and the linear model

The next proposition helps characterize the equivalence between the linear model (1) and the sequence model (22).

Proposition 10. *Suppose that $d \leq n$, $m = d$, and $\Sigma = (X^T X)^{-1}$.*

(a) If $\boldsymbol{\beta} = \boldsymbol{\theta}$ and $\mathbf{z} = (X^T X)^{-1} X^T \mathbf{y}$, then $\hat{\boldsymbol{\beta}}_{unif}(c) = \hat{\boldsymbol{\theta}}_{unif}(c)$, $\hat{\boldsymbol{\beta}}_r(c) = \hat{\boldsymbol{\theta}}_r(c)$, and

$$R\{\hat{\boldsymbol{\beta}}_r(c), \boldsymbol{\beta}\} = \tilde{R}\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\}.$$

(b) If $||\boldsymbol{\theta}|| = ||\boldsymbol{\beta}|| = c$, then

$$\begin{aligned} R\{\hat{\boldsymbol{\beta}}_{unif}(c), \boldsymbol{\beta}\} &= R^{(e)}(\boldsymbol{\beta}) = R^{(s)}(c) \\ &= \tilde{R}^{(s)}(c) = \tilde{R}^{(e)}(\boldsymbol{\theta}) = \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\}. \end{aligned}$$

Part (a) of Proposition 10 is obvious; part (b) follows from the fact that $((X^T X)^{-1} X^T \mathbf{y}, (X^T X)^{-1})$ is sufficient for $\boldsymbol{\beta}$ and the Rao-Blackwell inequality. Proposition 10 implies that it suffices to consider the sequence model in order to prove Theorem 1.

5. Proof of Theorem 1 (a): Normal approximation for the uniform prior

It follows from Proposition 9 that the Bayes estimator $\hat{\boldsymbol{\theta}}_{unif}(c)$ is optimal among all orthogonally equivariant estimators for $\boldsymbol{\theta}$, if $||\boldsymbol{\theta}|| = c$. In this section, we prove Theorem 1 (a) by bounding

$$\left| R\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\} - R\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\} \right| \tag{23}$$

and applying Proposition 10.

Marchand (1993) studied the relationship between $\hat{\boldsymbol{\theta}}_{unif}(c)$ and $\hat{\boldsymbol{\theta}}_r(c)$ under the assumption that $||\boldsymbol{\theta}|| = c$ and $\Sigma = \tau^2 I$ (i.e. in the Gaussian sequence model with iid errors). Marchand proved the following result, which is one of the keys to the proof of Theorem 1 (a).

Proposition 11 (Theorem 3.1 from (Marchand, 1993))**.** *Suppose that $\Sigma = \tau^2 I$ and $||\theta|| = c$. Then*

$$\begin{aligned}\left|\tilde{R}\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\} - \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\}\right| &\leq \frac{1}{m}\frac{c^2\tau^2 m}{c^2 + \tau^2 m} \\ &= \frac{1}{m}\tilde{R}\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\}.\end{aligned}$$

Thus, in the Gaussian sequence model with iid errors, the risk of $\hat{\boldsymbol{\theta}}_r(c)$ is nearly as small as that of $\hat{\boldsymbol{\theta}}_{unif}(c)$. Marchand's result relies on somewhat delicate calculations involving modified Bessel functions (Robert, 1990). A direct approach to bounding (23) for general Σ might involve attempting to mimic these calculations. However, this seems daunting (Bickel, 1981). Brown's identity, which relates the risk of a Bayes estimator to the Fisher, allows us to sidestep these calculations and apply Marchand's result directly.

Define the Fisher information of a random vector $\boldsymbol{\xi} \in \mathbb{R}^m$, with density $f_{\boldsymbol{\xi}}$ (with respect to Lebesgue measure on $\mathbb{R}^m$) by

$$I(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} \frac{\nabla f_{\boldsymbol{\xi}}(\mathbf{t})\nabla f_{\boldsymbol{\xi}}(\mathbf{t})^T}{f_{\boldsymbol{\xi}}(\mathbf{t})}\, d\mathbf{t},$$

where $\nabla f_{\boldsymbol{\xi}}(\mathbf{t})$ is the gradient of $f_{\boldsymbol{\xi}}(\mathbf{t})$. Brown's identity has typically been used for univariate problems or problems in the sequence model with iid Gaussian errors (Bickel, 1981; Brown and Gajek, 1990; Brown and Low, 1991; DasGupta, 2010). The next proposition is a straightforward generalization to the correlated multivariate Gaussian setting. Its proof is based on Stein's lemma.

Proposition 12 (Brown's Identity)**.** *Suppose that* $\mathrm{rank}(\Sigma) = m$*, with probability 1. Let $I_\Sigma(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}_0)$ denote the Fisher information of $\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}_0$, conditional on Σ, under the assumption that $\boldsymbol{\theta} \sim \pi_c$ is independent of $\boldsymbol{\delta}_0$ and Σ. If $||\boldsymbol{\theta}|| = c$, then*

$$\tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\} = E\mathrm{tr}(\Sigma) - E\mathrm{tr}\left\{\Sigma^2 I_\Sigma(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta})\right\}.$$

Proof. Suppose that $c = ||\boldsymbol{\theta}||$ and let

$$f(\mathbf{z}) = \int_{S(c)} (2\pi)^{-d/2}\det(\Sigma^{-1/2})e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\theta})^T\Sigma^{-1}(\mathbf{z}-\boldsymbol{\theta})}\, d\pi_c(\boldsymbol{\theta})$$

be the density of $\mathbf{z} = \boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}_0$, conditional on Σ and under the assumption that $\boldsymbol{\theta} \sim \pi_c$. Then

$$\hat{\boldsymbol{\theta}}_{unif}(c) = E_{\pi_c}(\boldsymbol{\theta}|\mathbf{z}, \Sigma) = \mathbf{z} - E_{\pi_c}(\Sigma^{1/2}\boldsymbol{\delta}_0|\mathbf{z}, \Sigma) = \mathbf{z} + \frac{\Sigma\nabla f(\mathbf{z})}{f(\mathbf{z})}.$$

It follows that

$$\begin{aligned} E||\hat{\boldsymbol{\theta}}_{unif}(c) - \boldsymbol{\theta}||^2 &= E\left|\left|\Sigma^{1/2}\boldsymbol{\delta} + \frac{\Sigma\nabla f(\mathbf{z})}{f(\mathbf{z})}\right|\right|^2 \\ &= E\mathrm{tr}(\Sigma) + 2E\left\{\frac{\boldsymbol{\delta}^T\Sigma^{3/2}\nabla f(\mathbf{z})}{f(\mathbf{z})}\right\} \\ &\qquad + E\left\{\frac{\nabla m(\mathbf{z})^T\Sigma^2\nabla f(\mathbf{z})}{f(\mathbf{z})^2}\right\} \\ &= E\mathrm{tr}(\Sigma) + 2E\left\{\frac{\boldsymbol{\delta}^T\Sigma^{3/2}\nabla f(\mathbf{z})}{f(\mathbf{z})}\right\} \\ &\qquad + E\mathrm{tr}\left\{\Sigma^2 I_\Sigma(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta})\right\} \end{aligned} \tag{24}$$

By Stein's lemma (integration by parts),

$$\begin{aligned} E\left\{\frac{\boldsymbol{\delta}^T\Sigma^{3/2}\nabla f(\mathbf{z})}{f(\mathbf{z})}\right\} &= E\left[\mathrm{tr}\left\{\Sigma^2\nabla^2\log f(\mathbf{z})\right\}\right] \\ &= -E\mathrm{tr}\left\{\Sigma^2 I_\Sigma(\boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta})\right\}. \end{aligned} \tag{25}$$

Brown's identity follows by combining (24) and (25). □

Using Brown's identity, Fisher information bounds may be converted to risk bounds, and vice-versa. Its usefulness in the present context springs from (i) the decomposition

$$\mathbf{z} = \boldsymbol{\theta} + \Sigma^{1/2}\boldsymbol{\delta}_0 = \left\{\boldsymbol{\theta} + (\gamma s_m)^{1/2}\boldsymbol{\delta}_1\right\} + (\Sigma - \gamma s_m)^{1/2}\boldsymbol{\delta}_2, \tag{26}$$

where $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \overset{\text{iid}}{\sim} N(0, I_m)$ are independent of Σ, s_m is the smallest eigenvalue of Σ, and $0 < \gamma < 1$ is a constant and (ii) Stam's inequality for the Fisher information of sums of independent random variables.

Proposition 13 (Stam's inequality; this version due to Zamir (1998))**.** *Let* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^m$ *be independent random variables that are absolutely continuous with respect to Lebesgue measure on* $\mathbb{R}^m$*. For every* $m \times m$ *positive definite matrix* Σ*,*

$$\mathrm{tr}\left[\Sigma^2 I(\mathbf{v} + \mathbf{w})\right] \leq \mathrm{tr}\left\{\Sigma^2\left[I(\mathbf{v})^{-1} + I(\mathbf{w})^{-1}\right]^{-1}\right\}.$$

Notice that conditional on Σ, the term $\boldsymbol{\theta} + (\gamma s_m)^{1/2}\boldsymbol{\delta}_1$ in (26) may be viewed as an observation from the Gaussian sequence model with iid errors. The necessary bound on (23) is obtained by piecing together Brown's identity, the decomposition (26), and Stam's inequality, so that Marchand's inequality (Proposition 11) may be applied to $\boldsymbol{\theta} + (\gamma s_m)^{1/2}\boldsymbol{\delta}_1$.

Proposition 14. *Suppose that Σ has rank m with probability 1 and that $||\boldsymbol{\theta}|| = c$. Let $s_1 \geq \cdots \geq s_m \geq 0$ denote the eigenvalues of Σ. Then*

$$\left|\tilde{R}\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\} - R\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\}\right| \leq \frac{1}{m} E\left\{\frac{s_1}{s_m} \mathrm{tr}\left(\Sigma^{-1} + m/c^2 I\right)^{-1}\right\}.$$

Proof. It is straightforward to check that

$$R\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\} = E\mathrm{tr}(\Sigma^{-1} + m/c^2 I)^{-1}. \tag{27}$$

Thus, Brown's identity and (27) imply

$$\begin{aligned}\tilde{R}\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\} - \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\} &= E\mathrm{tr}\left\{\Sigma^2 I_\Sigma(\boldsymbol{\theta} + \boldsymbol{\delta})\right\} \\ &\qquad + E\mathrm{tr}(\Sigma^{-1} + m/c^2 I)^{-1} - E\mathrm{tr}(\Sigma) \\ &= E\mathrm{tr}\left\{\Sigma^2 I_\Sigma(\boldsymbol{\theta} + \boldsymbol{\delta})\right\} \\ &\qquad - E\mathrm{tr}\left\{\Sigma^2(\Sigma + c^2/mI)^{-1}\right\}.\end{aligned}$$

Taking $\mathbf{v} = \boldsymbol{\theta} + (\gamma s_m)^{1/2}\boldsymbol{\delta}_1$ and $\mathbf{w} = (\Sigma - \gamma s_m)^{1/2}\boldsymbol{\delta}_2$ in Stam's inequality, where $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$, and $0 < \gamma < 1$ are given in (26), one obtains

$$\begin{aligned}\tilde{R}\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\} - \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\} &\leq E\mathrm{tr}\left(\Sigma^2\Big[I_\Sigma\{\boldsymbol{\theta} + (\gamma s_m)^{1/2}\boldsymbol{\delta}_1\}^{-1}\right. \\ &\qquad \left. + \Sigma - \gamma s_m I\Big]^{-1}\right) \\ &\qquad - E\mathrm{tr}\left\{\Sigma^2(\Sigma + c^2/mI)^{-1}\right\}\end{aligned}$$

By orthogonal invariance, $I_\Sigma\{\boldsymbol{\theta} + (\gamma s_m)^{1/2}\boldsymbol{\delta}_1\} = \zeta I_m$ for some $\zeta \geq 0$. Marchand's inequality, another application of Brown's identity, and (27) with $\Sigma = \gamma s_m I_m$ imply that

$$\zeta \leq \left(\frac{1}{\gamma s_m}\right)\frac{\gamma s_m + c^2/m^2}{\gamma s_m + c^2/m}.$$

Since

$$\frac{1}{\zeta} - \gamma s_m \geq (m-1)\frac{\gamma s_m c^2}{\gamma s_m m^2 + c^2},$$

it follows that

$$\begin{aligned}\tilde{R}\{\hat{\boldsymbol{\theta}}_r(c), \boldsymbol{\theta}\} - \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c), \boldsymbol{\theta}\} &\leq E\mathrm{tr}\left[\Sigma^2\left\{\Sigma + (m-1)\frac{\gamma s_m c^2}{\gamma s_m m^2 + c^2} I\right\}^{-1}\right] \\ &\qquad - E\mathrm{tr}\left\{\Sigma^2(\Sigma + c^2/mI)^{-1}\right\}.\end{aligned}$$

Taking $\gamma \uparrow 1$,

$$\begin{aligned}\tilde{R}\{\hat{\boldsymbol{\theta}}_r(c),\boldsymbol{\theta}\} - \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c),\boldsymbol{\theta}\} &\leq E\mathrm{tr}\left[\Sigma^2\left\{\Sigma + (m-1)\frac{s_m c^2}{s_m m^2 + c^2}I\right\}^{-1}\right] \\ &\quad -E\mathrm{tr}\left\{\Sigma^2(\Sigma + c^2/mI)^{-1}\right\} \\ &\leq \frac{1}{m}E\left\{\frac{s_1}{s_m}\mathrm{tr}\left(\Sigma^{-1} + m/c^2 I\right)^{-1}\right\}.\end{aligned}$$

The proposition follows because $\tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(c),\boldsymbol{\theta}\} \leq \tilde{R}\{\hat{\boldsymbol{\theta}}_r(c),\boldsymbol{\theta}\}$. □

Theorem 1 (a) follows immediately from Propositions 10 and 14.

6. Proof of Theorem 1 (b): $d > n$

It only remains to prove Theorem 1 (b), which is achieved through a sequence of lemmas. The first step of the proof focuses on the linear model (as opposed to the sequence model) and on reducing the problem where $d > n$ and X^TX is not invertible to a full rank problem. This step builds on Lemma 1 from Section 2.4.

Suppose that $d > n$ and let $X = UDV^T$ be the singular value decomposition of X, where $U \in O(n)$, $V \in O(d)$, $D = (D_0 \;\; 0)$, and D_0 is a rank n diagonal matrix (with probability 1). Let $W \in O(d)$ be uniformly distributed on $O(n)$ (according to Haar measure) and independent of $\boldsymbol{\epsilon}$ and X. Define the $n \times n$ matrix $X_0 = UD_0W^T$ and consider the full rank linear model

$$\mathbf{y}_0 = X_0\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \tag{28}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^n$. Notice that unlike X, the entries in X_0 are *not* iid $N(0,1)$. However, $X_0^TX_0$ is orthogonally invariant. As with the linear model (1), one can consider estimators $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_0(\mathbf{y}_0, X_0)$ for $\boldsymbol{\beta}_0$ and compute the risk

$$R_0(\hat{\boldsymbol{\beta}}_0, \boldsymbol{\beta}_0) = E_{\boldsymbol{\beta}_0}||\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0||^2, \tag{29}$$

where the expectation in (29) is taken over $\boldsymbol{\epsilon}$ and X_0. We have the following lemma.

Lemma 2. *Suppose that $d > n$, $||\boldsymbol{\beta}|| = c$, and $\hat{\boldsymbol{\beta}} \in \mathscr{E}(n,d)$. Let P_0 denote any fixed $n \times d$ projection matrix with orthogonal rows. Then there is an orthogonally equivariant estimator $\mathscr{P}_0\hat{\boldsymbol{\beta}} \in \mathscr{E}(n,n)$ such that*

$$R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \int_{S_d(c)} R_0(\mathscr{P}_0\hat{\boldsymbol{\beta}}, P_0\mathbf{b})\; d\pi_c(\mathbf{b}) + \frac{d-n}{d}c^2.$$

Proof. As above, let $X = UDV^T$ be the singular value decomposition of X. Let V_0 denote the first n columns of V and let V_1 denote the remaining $d-n$ columns of V. By (8),

$$\hat{\boldsymbol{\beta}}(\mathbf{y}, X) = V_0\hat{\boldsymbol{\beta}}_0(\mathbf{y}, UD_0),$$

where $\mathscr{P}_0\hat{\boldsymbol{\beta}}(\mathbf{y}, UD_0) = \hat{\boldsymbol{\beta}}_0(\mathbf{y}, UD_0)$ is the first n coordinates of $\hat{\boldsymbol{\beta}}(\mathbf{y}, UD)$. Furthermore, it is easy to check that $\mathscr{P}_0\hat{\boldsymbol{\beta}}$ is orthogonally equivariant, i.e. $\mathscr{P}_0\hat{\boldsymbol{\beta}} \in \mathscr{E}(n,n)$. Thus,

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= E_{\boldsymbol{\beta}}||\hat{\boldsymbol{\beta}}_0(\mathbf{y}, UD_0) - V_0^T\boldsymbol{\beta}||^2 + E_{\boldsymbol{\beta}}||V_1^T\boldsymbol{\beta}||^2 \\ &= E_{\boldsymbol{\beta}}||\hat{\boldsymbol{\beta}}_0(\mathbf{y}, UD_0) - V_0^T\boldsymbol{\beta}||^2 + \frac{d-n}{d}c^2. \end{aligned}$$

To prove the lemma, it suffices to show that

$$E_{\boldsymbol{\beta}}||\hat{\boldsymbol{\beta}}_0(\mathbf{y}, UD_0) - V_0^T\boldsymbol{\beta}||^2 = \int_{S_d(c)} R_0(\hat{\boldsymbol{\beta}}_0, P_0\mathbf{b})\ d\pi_c(\mathbf{b}).$$

By Proposition 1, orthogonal invariance of π_c, and orthogonal equivariance of $\hat{\boldsymbol{\beta}}_0$,

$$\begin{aligned} E_{\boldsymbol{\beta}}||\hat{\boldsymbol{\beta}}_0(\mathbf{y}, UD_0) - V_0^T\boldsymbol{\beta}||^2 &= \int_{S_d(c)} E_{\mathbf{b}}||\hat{\boldsymbol{\beta}}_0(\mathbf{y}, UD_0) - V_0^T\mathbf{b}||^2\ d\pi_c(\mathbf{b}) \\ &= E\left\{\int_{S_d(c)} ||\hat{\boldsymbol{\beta}}_0(UD_0V_0^T\mathbf{b} + \boldsymbol{\epsilon}, UD_0) \right. \\ &\qquad\qquad \left. - V_0^T\mathbf{b}||^2\ d\pi_c(\mathbf{b})\right\} \\ &= E\left\{\int_{S_d(c)} ||\hat{\boldsymbol{\beta}}_0(UD_0W^TP_0\mathbf{b} + \boldsymbol{\epsilon}, UD_0) \right. \\ &\qquad\qquad \left. - W^TP_0\mathbf{b}||^2\ d\pi_c(\mathbf{b})\right\} \\ &= \int_{S_d(c)} E||\hat{\boldsymbol{\beta}}_0(\mathbf{y}_0, X_0) - P_0\mathbf{b}||^2\ d\pi_c(\mathbf{b}), \end{aligned}$$

as was to be shown. □

Lemma 2 allows us to express the risk of an equivariant estimator for $\boldsymbol{\beta}$ in the linear model (1) with $d > n$ in terms of the risk of another equivariant estimator in a different linear model (28) with $d = n$. Though the linear model (28) differs from the original linear model with Guassian predictors – thus, Theorem 1 (a) does not apply directly – (28) is equivalent to the sequence model (22), with $m = n$ and $\Sigma = (X_0^TX_0)^{-1}$.

Lemma 3. *Suppose that $2 < m = n < d$ and $\Sigma = (X_0^T X_0)^{-1}$ in the sequence model (22). Also suppose that $||\boldsymbol{\beta}|| = c$. Let P_0 be a fixed $n \times d$ projection matrix with orthogonal rows and let $s_1 \geq \cdots \geq s_n \geq 0$ denote the eigenvalues of $(X^T X)^{-1}$. Then*

$$\begin{aligned} R\{\hat{\boldsymbol{\beta}}_{unif}(c), \boldsymbol{\beta}\} &\geq \int_{S_d(c)} \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(P_0\mathbf{t}), P_0\mathbf{t}\}\ d\pi_c(\mathbf{t}) + \frac{d-n}{d}c^2 \\ &\geq \int_{S_d(c)} E\left\{\left(1 - \frac{s_1}{ns_n}\right) \mathrm{tr}\left(XX^T + \frac{n}{||P_0\mathbf{t}||^2}I\right)^{-1}\right\} d\pi_c(\mathbf{t}) \\ &\qquad + \frac{d-n}{d}c^2 \\ &\geq E\left[\left(1 - \frac{s_1}{ns_n}\right) \mathrm{tr}\left\{XX^T + \frac{n(d-2)}{c^2(n-2)}I\right\}^{-1}\right] + \frac{d-n}{d}c^2. \end{aligned}$$

Proof. The first inequality follows from Lemma 2 and a suitably modified version of Proposition 10 that describes the equivalence between the linear model (28) and the sequence model (22). The second inequality follows from Proposition 14 and the fact that $X_0^T X_0$ and XX^T have the same eigenvalues:

$$\begin{aligned} \tilde{R}\{\hat{\boldsymbol{\theta}}_{unif}(P_0\mathbf{t}), P_0\mathbf{t}\} &\geq \tilde{R}\{\hat{\boldsymbol{\theta}}_r(P_0\mathbf{t}), P_0\mathbf{t}\} \\ &\qquad - \frac{1}{n}E\left\{\frac{s_1}{s_n}\mathrm{tr}(X_0X_0^T + n/||P_0\mathbf{t}||^2 I)^{-1}\right\} \\ &= E\left\{\left(1 - \frac{s_1}{ns_n}\right)\mathrm{tr}\left(X_0^T X_0 + n/||P_0\mathbf{t}||^2 I\right)^{-1}\right\} \\ &= E\left\{\left(1 - \frac{s_1}{ns_n}\right)\mathrm{tr}\left(XX^T + n/||P_0\mathbf{t}||^2 I\right)^{-1}\right\}. \end{aligned}$$

The last inequality in the lemma follows from Jensen's inequality and the identity

$$\int_{S_d(c)} \frac{1}{||P_0\mathbf{t}||^2}\ d\pi_c(\mathbf{t}) = \frac{d-2}{c^2(n-2)}.$$

□

We now have the tools to complete the proof of Theorem 1 (b). Suppose that $d > n$ and $||\boldsymbol{\beta}|| = c$. Then

$$R(\hat{\boldsymbol{\beta}}_r^*, \boldsymbol{\beta}) = E\mathrm{tr}\{XX^T + d/c^2 I\}^{-1} + \frac{d-n}{d}c^2.$$

Since $R\{\hat{\boldsymbol{\beta}}_r(c),\boldsymbol{\beta}\} - R\{\hat{\boldsymbol{\beta}}_{unif}(c),\boldsymbol{\beta}\} = R\{\hat{\boldsymbol{\beta}}_r(c),\boldsymbol{\beta}\} - R^{(e)}(\boldsymbol{\beta}) \geq 0$, Lemma 3 implies

$$\begin{aligned}
\left|R\{\hat{\boldsymbol{\beta}}_r(c),\boldsymbol{\beta}\} - R^{(e)}(\boldsymbol{\beta})\right| &\leq E\mathrm{tr}\{XX^T + d/c^2I\}^{-1} \\
&\quad -E\left[\left(1-\frac{s_1}{ns_n}\right)\mathrm{tr}\left\{XX^T + \frac{n(d-2)}{c^2(n-2)}I\right\}^{-1}\right] \\
&\leq \frac{1}{n}E\left\{\frac{s_1}{s_n}\mathrm{tr}(XX^T + d/c^2I)^{-1}\right\} \\
&\quad\quad +2\frac{d-n}{c^2(n-2)}E\mathrm{tr}(XX^T + d/c^2I)^{-2}.
\end{aligned}$$

Theorem 1 (b) follows.

Acknowledgements

The author thanks Sihai Zhao for his thoughtful comments and suggestions.

References

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics* **34** 584–653.

Bai, Z. (1993). Convergence rate of expected spectral distributions of large random matrices. Part II. Sample covariance matrices. *Annals of Probability* **21** 649–672.

Bansal, V., Libiger, O., Torkamani, A. and Schork, N. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics* **11** 773–785.

Baranchik, A. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Annals of Statistics* **1** 312–321.

Beran, R. (1996). Stein estimation in high dimensions: A retrospective. In *Research developments in probability and statistics: Festschrift in honor of Madan L. Puri on the occasion of his 65th birthday*. VSP International Science Publishers.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer.

Bickel, P. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Annals of Statistics* **9** 1301–1309.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics* **37** 1705–1732.

Borel, É. (1914). *Introduction géométrique à quelques théories physiques*. Gauthier-Villars.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association* **78** 131–136.

Brown, L. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics* **42** 855–903.

Brown, L. (1990). An ancillarity paradox which appears in multiple linear regression. *Annals of Statistics* **18** 471–493.

Brown, L. and Gajek, L. (1990). Information inequalities for the bayes risk. *Annals of Statistics* **18** 1578–1594.

Brown, L. and Low, M. (1991). Information inequality bounds on the minimax risk (with an application to nonparametric regression). *Annals of Statistics* **19** 329–337.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* **1** 169–194.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351.

Cavalier, L. and Tsybakov, A. (2002). Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields* **123** 323–354.

DasGupta, A. (2010). False vs. missed discoveries, gaussian decision theory, and the donsker-varadhan principle. In *Borrowing Strength: Theory Powering Applications A Festschrift for Lawrence D. Brown*. Institute of Mathematical Statistics.

Diaconis, P. and Freedman, D. (1987). A dozen de finetti-style results in search of a theory. *Annales de l'Henri Poincaré, Probabilités et Statistiques* **23** 397–423.

Dicker, L. (2012). Dense signals, linear estimators, and out-of-sample prediction for high-dimensional linear models. Preprint.

Donoho, D. (1995). De-noising by soft-thresholding. *Information Theory, IEEE Transactions on* **41** 613–627.

Donoho, D. and Johnstone, I. (1994). Minimax risk over ℓ^p-balls for ℓ^q-error. *Probability Theory and Related Fields* **99** 277–303.

Duarte, M., Davenport, M., Takhar, D., Laska, J., Sun, T., Kelly, K. and Baraniuk, R. (2008). Single-pixel imaging via compressive sampling. *Signal Processing Magazine, IEEE* **25** 83–91.

Erlich, Y., Gordon, A., Brand, M., Hannon, G. and Mitra, P. (2010). Compressed genotyping. *Information Theory, IEEE Transactions on* **56** 706–723.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 37–65.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on* **57** 5467–5484.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004). Discussion of boosting papers. *Ann. Statist* **32** 102–107.

Goldenshluger, A. and Tsybakov, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *Annals of Statistics* **29** 1601–1619.

Goldenshluger, A. and Tsybakov, A. (2003). Optimal prediction for linear regression with infinitely many parameters. *Journal of Multivariate Analysis* **84** 40–60.

Goldstein, D. (2009). Common genetic variation and human traits. *New England Journal of Medicine* **360** 1696–1698.

Hall, P., Jin, J. and Miller, H. (2009). Feature selection when there are many influential features. Arxiv preprint arXiv:0911.4076.

Hirschhorn, J. (2009). Genomewide association studies – illuminating biologic pathways. *New England Journal of Medicine* **360** 1699–1701.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: held at the Statistical Laboratory, University of California, June 20-July 30, 1960*. University of California Press.

Johnstone, I. (2011). Gaussian Estimation: Sequence and Wavelet Models. Unpublished manuscript.

Kraft, P. and Hunter, D. (2009). Genetic risk prediction – are we there yet? *New England Journal of Medicine* **360** 1701–1703.

Leeb, H. (2009). Conditional predictive inference post model selection. *Annals of Statistics* **37** 2838–2876.

Lévy, P. (1922). *Leçons d'Analyse Fonctionnelle*. Gauthier-Villars.

Lustig, M., Donoho, D. and Pauly, J. (2007). Sparse MRI: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine* **58** 1182–1195.

Manolio, T. (2010). Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine* **363** 166–176.

Marčenko, V. and Pastur, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR–Sbornik* **1** 457–483.

Marchand, E. (1993). Estimation of a multivariate mean with constraints on the norm. *Canadian Journal of Statistics* **21** 359–366.

Pinsker, M. (1980). Optimal filtration of functions from l2 in gaussian noise. *Problems of Information Transmission* **16** 52–68.

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Annals of Statistics* **39** 731–771.

Robert, C. (1990). Modified bessel functions and their applications in probability and statistics. *Statistics & probability letters* **9** 155–161.

Stam, A. (1959). Some inequalities satisfied by the quantities of information of fisher and shannon1. *Information and Control* **2** 101–112.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, vol. 1.

Stein, C. (1960). Multiple regression. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press.

Sun, T. and Zhang, C. (2011). Scaled sparse linear regression. Arxiv preprint arXiv:1104.4595.

Tikhonov, A. (1943). On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* **39** 195–198.

Wright, J., Yang, A., Ganesh, A., Sastry, S. and Ma, Y. (2008). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31** 210–227.

Zamir, R. (1998). A proof of the fisher information inequality via a data processing argument. *Information Theory, IEEE Transactions on* **44** 1246–1250.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–942.